CENG463-Programming Assignment 2

Cansu Eskici

1. Introduction

For this assignment, I chose Latvia as the country. For fine-tuning, I chose multilingual BERT as my model. For the multilingual casual language model, I chose Llama-3.1-8B. I worked on Google Colab with A100 GPU. This project can be found in this Github repository.

2. Task 1

2.1. Approach

For this task, I used pandas, sklearn, transformers, and torch libraries. After reading the orientation data into a variable, I wanted to see how they were distributed. I used the value_counts() method of pandas to see the distribution of the data, and it turned out to be quite imbalanced. There were 628 data instances for label 1, and 170 for label 0. For splitting the data into train and validation datasets, I used train_test_split from sklearn.model_selection. %90 of data was split into training, while %10 went into validation. Distribution of both of the datasets were still imbalanced.

As mentioned above, multilingual BERT was chosen to be the multilingual masked language model since it is specifically designed to handle multiple languages. The tokenizer and model was initialized. For tokenization, a tokenize function was defined to preprocess the text data. Both of the datasets were converted from pandas DataFrames to HuggingFace Datasets, and tokenize was mapped to them. By setting up TrainingArguments and Trainer, the model was fine-tuned. The function, compute_metrics is defined to evaluate the model.

For casual language model, Llama-3.1-8B was chosen. The steps were similar to the multilingual model. The validation data was loaded and preprocessed by the tokenizer. The classify_batch function was defined to process batches of data, and generate predictions according to the prompt. From the library sklearn.metrics, classification_report was used to evaluate the model.

CENG463-Programming Assignment 2

CEUR Workshop Proceedings (CEUR-WS.org)

2.2. Results

2.2.1. Results for classifying in original language.

Table 1
Results with Multilingual BERT

	Precision	Recall	F1-Score
Class 0	0.00	0.00	0.80
Class 1	0.79	1.00	0.88
Accuracy			0.79
Macro Avg	0.39	0.50	0.44
Weighted Avg	0.62	0.79	0.69

Table 2
Results with Llama-3.1-8B

	Precision	Recall	F1-Score
Class 0	0.24	0.72	0.36
Class 1	0.84	0.39	0.53
Accuracy			0.46
Macro Avg	0.54	0.55	0.45
Weighted Avg	0.71	0.46	0.50

2.2.2. Results for classifying in English.

Table 3
Results with Multilingual BERT

	Precision	Recall	F1-Score
Class 0	0.73	0.47	0.57
Class 1	0.87	0.95	0.91
Accuracy			0.85
Macro Avg	0.80	0.71	0.74
Weighted Avg	0.84	0.85	0.84

Table 4
Results with Llama-3.1-8B

	Precision	Recall	F1-Score
Class 0	0.23	0.74	0.35
Class 1	0.82	0.32	0.46
Accuracy			0.41
Macro Avg	0.53	0.53	0.41
Weighted Avg	0.70	0.41	0.44

2.3. Discussion

Multilingual BERT has higher accuracy than Llama-3.1-8B in both original language and English classifications. The big margin between the precision and recall values for Multilingual BERT and original language stems from imbalance in the dataset, and the lack of training of the model in that language. For the same dataset, using english data results in better performance scores. The results indicate that Multilingual BERT is more effective for this classification task compared to Llama-3.1-8.

3. Task 2

3.1. Approach

For this task, the same implementations were used. The only difference was the dataset, instead of using orientation dataset, power dataset were used. Also for the causal language model, the prompt was changed in order to classify if the speech is in favor of opposition or governing.

3.2. Results

3.2.1. Results for classifying in original language.

Table 5
Results with Multilingual Bert

	Precision	Recall	F1-Score
Class 0	0.69	0.94	0.79
Class 1	0.54	0.15	0.23
Accuracy			0.67
Macro Avg	0.61	0.54	0.51
Weighted Avg	0.64	0.67	0.61

Table 6
Results with Llama-3.1-8B

	Precision	Recall	F1-Score
0	0.75	0.27	0.39
1	0.35	0.82	0.49
Accuracy			0.45
Macro Avg	0.55	0.54	0.44
Weighted Avg	0.62	0.45	0.43

3.2.2. Results for classifying in English.

Table 7
Results with Multilingual Bert

	Precision	Recall	F1-Score
Class 0	0.69	0.82	0.75
Class 1	0.43	0.28	0.34
Accuracy			0.64
Macro Avg	0.56	0.55	0.54
Weighted Avg	0.61	0.64	0.61

Table 8
Results with Llama-3.1-8B

	Precision	Recall	F1-Score
0	0.72	0.57	0.64
1	0.39	0.56	0.46
Accuracy			0.57
Macro Avg	0.56	0.57	0.55
Weighted Avg	0.61	0.57	0.58

3.3. Discussion

For precision and recall, models show similar results. For this task, multilingual BERT slightly works better with the original language than English. For Llama-3.1-8B, the results are similar for both languages. Multilingual BERT shows higher precision, recall, and F1-scores for Class 0 in both original language and English classifications. Multilingual BERT has a higher overall accuracy compared to Llama-3.1-8B in both original language and English classifications. The results indicate that Multilingual BERT is more effective for this classification task compared to Llama-3.1-8B, again.